REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
30-06-2006	Final	Dec 2001 - Mar 2006

4. TITLE AND SUBTITLE
An Architecture for a Problem-Solving Assessment Authoring and Delivery System

5a. CONTRACT NUMBER

5b. GRANT NUMBER
N00014-02-1-0179

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Gregory K. W. K. Chung, Eva L. Baker, Girlie C. Delacruz, Jesse J. Elmore, William L. Bewley, and Bruce Seely

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
UCLA CSE/CRESST
300 Charles E. Young Dr. North
300 GSE&IS/Mailbox 951522
Los Angeles, CA 90095-1522

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)
ONR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A
Approved for Public Release
Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The UCLA National Center for Research on Evaluation, Standards, and Student Testing (CRESST) designed an authoring system to support the design of problem-solving assessments. To support assessment design, the system design includes a constraint network, in which nodes are variables that can assume a range of values and the topology specifies how the variables and values are related, to describe the permissible relations and states among assessment and problem-solving variables. This report describes the design for the authoring system and discusses next steps in developing and conducting research on the system.

15. SUBJECT TERMS
problem solving, assessment design, assessment authoring

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
Unclassified	Unclassified	Unclassified			19b. TELEPHONE NUMBER *(Include area code)*

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

An Architecture for a Problem-Solving Assessment Authoring and Delivery System

Deliverable – June 2006

Knowledge, Models and Tools to Improve the Effectiveness of Naval Distance Learning

Eva L. Baker
CRESST/University of California, Los Angeles

Office of Naval Research
Award # N00014-02-1-0179

National Center for Research on Evaluation, Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
301 GSE&IS, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under Office of Naval Research Award Number #N00014-02-1-0179, as administered by the Office of Naval Research. The findings and opinions expressed in this report do not necessarily reflect the positions or policies of the Office of Naval Research.

TABLE OF CONTENTS

AN ARCHITECTURE FOR A PROBLEM-SOLVING ASSESSMENT AUTHORING AND DELIVERY SYSTEM

Gregory K. W. K. Chung, Eva L. Baker, Girlie C. Delacruz, Jesse J. Elmore,

William L. Bewley, and Bruce Seely

CRESST/University of California, Los Angeles

Abstract

This report describes the design of an authoring system to support the design of problem-solving assessments. A key component underlying the system architecture is a constraint network. In a constraint network, nodes are variables that can assume a range of values and the topology specifies how the variables and values are related (Montanari, 1974). To support assessment design, the system design includes a constraint network describing the permissible relations and states among assessment and problem-solving variables.

The UCLA National Center for Research on Evaluation, Standards, and Student Testing (CRESST) is under contract to the Office of Naval Research (ONR) to conduct research on assessment models and tools designed to support Navy and Marine Corps distance learning (DL). The project is called Knowledge, Models, and Tools to Improve the Effectiveness of Naval Distance Learning, or KMT. The approach to conducting KMT research has been to develop and test tools designed to address the assessment and training requirements posed by real Navy and Marine Corps training applications.

The Navy is currently undergoing a revolution in training. Future training, as envisioned by Admiral Clark, "will apply information-age methods to accelerate learning and improve proficiency" (Clark, 2003). These methods include advanced trainers and simulators, tailored-skills training programs, and more effective performance measurement tools. Distributed learning is a fundamental element of the envisioned information-age methods, in part because of its promise to provide efficient and effective learning on demand, and in part because of its promise to reduce the time Sailors and Marines spend away from their command attending schools. The Navy is taking aggressive steps to support distance learning and the Revolution in Training by developing an Integrated Learning Environment (ILE), the objectives of which include reducing content development cost and lead time through reuse of learning objects and improving content relevance by reducing cycle times to review and validate requirements (NPDC, 2003).

To deliver these capabilities, the new Navy training organizations and training development contractors need research-based human performance and assessment knowledge, models, and tools to support the development and use of information-age methods. These include guidelines for courseware development and evaluation and courseware-authoring capabilities including models and tools for assessing learner performance based on pedagogically sound principles consistent with the best available knowledge from learning and assessment research and instructional design.

Authoring Systems for Assessment

In his reflection on the current state of testing, van der Linden (2005) observed,

> Any outsider entering the testing industry would expect to find a spin-off in the form of a well-developed technology that enables us to engineer tests rigorously to our

> specifications.... To draw a parallel with the natural sciences, it seems as if testing has led to the development of a new science, but the spin-off in the form of a technology for engineering the test has not yet been realized. (p. xi)

van der Linden's (2005) observation highlights the craft nature of assessment development: the lack of a common knowledge will prevent the field from moving from craft knowledge toward an engineering model for test development. What is needed is a method to explicitly represent the assessment design, from which actual assessments can be derived. Establishing this assessment model, in computational form, will provide traceability between the assessment (instantiated in a particular content area for a particular purpose and population) and the assessment model. Thus, the development of assessments becomes rational—particular features (good or bad) of the assessment are based on an underlying model. Because the model is in computational form, it is persistent until modified. This transparency is a prerequisite for demystifying the test development process and is central to moving toward a "technology for engineering the test."

One of the most important capabilities of an assessment authoring system would be the shortening in the time required to gather validity evidence for different purposes (Baker, 2002a; O'Neil & Baker, 1997). Historically, the development lifecycle of assessments, particularly for measuring complex learning (e.g., performance assessments), requires significant amounts of time and resources (O'Neil & Baker, 1997). Thus, an authoring system should have the capability to rapidly generate tasks appropriate for different assessment purposes.

When authoring is applied to the field of testing, additional requirements come into play. With assessment and testing, the key requirement is validity, that is, the extent to which inferences drawn from the result of the test or assessment are warranted (Messick, 1995). Linn, Baker, and Dunbar (1991) have described essential elements of validity applied to open-ended assessment tasks. These validity criteria include cognitive complexity, linguistic appropriateness, transfer and generalizability, content quality, reliability, and instructional sensitivity. Moreover, when designing an authoring system (rather than a test, for example) one is interested in the utility of the system for its users (teachers or test developers) in addition to the value of the data yielded by administering tests to students.

In testing, it is often the case that instructors who need to use tests and assessments routinely in their classrooms have little time and expertise to create high-quality assessments of student learning. They may use a craft approach, creating each test, one at a time, with a wholly new format, scoring approach, and set of cognitive requirements. This approach generally produces tests of low quality whose inferences may be suspect. As instructors attempt to bring all students up to high challenging standards, there is a concomitant desire to test performance in such a way as to stimulate complex cognitive processing. Usually, instructors use essay or other extended written examinations to elicit such performance. These types of tests are difficult to calibrate, take considerable time and cost to evaluate, and frequently result in low reliability in scoring. A straightforward way is needed for instructors to create assessments that require students to demonstrate complex knowledge representations, including declarative, procedural, and systemic knowledge. In this report we describe such an approach.

Prior and Current Assessment Authoring Systems

Prior work related to assessment authoring systems is anchored at three points. At one end are numerous resource Web sites that simply house existing assessments, whose quality is unknown. PALS (Quellmalz, Hinojosa, Hinojosa, & Schank, 2000) has been one of the few efforts to adopt a systematic vetting procedure to ensure that the assessments housed at their site meet a set of standards related to validity and reliability, but even so, using the assessments appropriately requires a relatively high level of assessment knowledge. More commonly, assessment sites are primarily warehouses for tasks with undefined technical properties (e.g., Eisenhower National Clearinghouse, 2001; Harvard Graduate School of Education, 2001; Queensland Association of Mathematics Teachers, 1998). Search and filter capabilities are usually provided to make it possible to retrieve assessments by different criteria (e.g., grade, task type, domain, standard). At the second point are systems that provide the means to build (but not design) assessments. These systems typically exist within learning management systems that have the infrastructure to deliver multiple test formats (e.g., multiple choice, true-false, short- and long-essay). However, the assessments are essentially forms that need to be filled in (e.g., CAPA, 2000; Learning Manager, 2005; QuestionMark, 2001; TRIADS, 2001; WebCT, 2001). At the third point are authoring systems that are targeted to assessment developers (Chung, Baker, & Cheak, 2001; Chung, Klein, Herl, & Bewley, 2001; Mislevy, Steinberg, Breyer, & Almond, 1999; Osmundson, Jeffries, & Herman, 1998). These systems build in some underlying knowledge about valid assessment design (e.g., concepts of validity and reliability) and may require detailed knowledge of the essential content. Other teacher-focused systems such as the Assessment Wizard (Educational Testing Service, 2000) expect less assessment knowledge, but they assume substantial domain knowledge.

More recent efforts by CRESST and others (e.g., the PADI project) have addressed the issue directly. For example, the ADDS system (Assessment Design and Delivery System), developed by CRESST, is intended to support teachers' creation of assessments mapping to standards, to challenging cognitive demands, and to subject matter content. Wizard interfaces guide teachers in the specification or selection of critical assessment criteria, and enable teachers to design their own assessments by graphical composition.

An important rationale underlying this work is the acknowledgment that in practical settings, finding relevant science content and examples for use in assessment is a time-consuming bottleneck for all assessment development. As a practical matter, teachers have neither the time nor expertise to find content for every assessment. Performance assessments or other memorable extended tasks are especially problematic because of potential learning effects—that is, once students engage in an assessment, they may learn important elements of the example used in the task or even the task itself.

The rationale for our approach is based on more than a decade of research and evaluation on the use of technology in classrooms (Baker, 2001, 2002b; Baker, Gearhart, & Herman, 1994; Baker & Herman, 2000; Baker, Herman, & Gearhart, 1996; Baker & O'Neil, 2003; Chen, Chung, Klein, de Vries, & Burnam, 2000; Chung, Klein, & Baker, 2000; Gearhart, Herman, Baker, Novak, & Whittaker, 1994; O'Neil & Baker, 1993). We have consistently found that teachers' adoption of technology is based largely on the perceived utility to improve students' learning, and the amount of overhead in time and effort in the non-instructional uses (e.g., equipment setup, loading software, amount of technical support needed). Our design choices directly address these issues by minimizing the demands associated with technology barriers (e.g., standard Web access and user-friendly interfaces) and knowledge and effort barriers (e.g., provision of default choices, automated access to content) while still providing the flexibility for teachers to design assessments for their particular needs.

Research Questions

An implicit assumption underlying the Navy's new training doctrine is the availability of assessments capable of measuring and providing quality information on trainees' knowledge, skills, and abilities. However, in practice, this assumption may not hold up for a variety of practical reasons (e.g., time constraints, not trained in assessment, and so forth). Thus, our research effort has been focused on developing methods to support non-assessment experts in developing assessments consistent with good design practices. The underlying assumption is that users will not be experts in designing assessments, and therefore the system should provide support that maximizes the likelihood of the assessments conforming to modern assessment practices. Thus, the set of research questions guiding this work are:

- To what extent can a problem-solving assessment framework be codified in computational form? What is an appropriate technique that is feasible and transparent?
- To what extent can a user's design of a problem-solving assessment be evaluated computationally? This is an important question for practical settings in which the users are assumed to have little assessment design expertise. Theoretically, establishing a framework from which assessments can be designed and evaluated against offers the potential for coherency, generalizability, and scalability.
- To what extent can qualitatively different assessment tasks be generated computationally? This question addresses issues of rapid deployment and scaling.

Problem-Solving Assessment Framework

The assessment framework used in this work is based on Baker and O'Neil's (2002) approach to designing problem-solving assessments using computer technology. This approach first characterizes three types of problem-solving tasks: (a) a task in which an appropriate solution is known in advance, (b) a task in which there is no known solution to the problem, and (c) a task that requires an application of a given tool set to a broadly ranging set of topics. Baker and O'Neil highlight the relevant variables that characterize these types of problems while focusing on complex, scenario-based problem-solving tasks. Table 1 shows the variables, possible values, and definitions for each variable and value.

Identifying the problem is often one of the most difficult aspects of problem solving (see Baker & O'Neil, 2002). The ambiguity of problem identification may be dependent upon the prior knowledge that is required of the examinee, as well as the purpose of the assessment. An assessment author could adjust its difficulty by either stating the problem explicitly or obscuring it in an embedded setting. Likewise, the information sources that make up the scenario can vary in quality from their accuracy to their credibility.

The problem to be solved can also be multiply-masked in which the solution to one part of the task determines the nature of the next part of the task. An examinee may be required to mentally test various hypotheses for solving the problem under extreme time constraints (e.g., an enemy is about to attack but you are not certain whether they are armed with rifles or tanks, or a house is on fire and you are unsure where the source of the fire is) and demonstrate proficiency in contingency planning in order to recover from an error (e.g., selected plan of attack underestimated the enemy's weapon power, or fire extinguisher does not work).

Baker and O'Neil (2002) emphasize the advantage of problems that are sequential and conditional in nature both because of their reflection of reality and their potential to measure competence. Computerized assessments take advantage of this ability to measure an examinee's proficiency of either the task outcome or on parts of the task. In this environment, an examinee can execute an action and either be given permission to continue, be given the opportunity to correct an invalid procedure, or be provided with a partial solution in order to proceed with the rest of the problem. By doing this, one can get a measure of an examinee's competence and incompetence on the relevant skills throughout the task. This affords the instructor the capacity to tailor instruction to that with which the examinee has trouble, rather than to instruct on every aspect of the task.

Table 1

Major Variables Affecting the Design of Problem-Solving Assessment Tasks

Variable (and values)	Definitions
ASSESSMENT PURPOSE (diagnostic, certification)	
COGNITIVE DEMAND (self-regulation, reasoning, content understanding, factual knowledge)	
PROBLEM IDENTIFICATION	Characteristics related to the nature of the problem. The author can modulate the Problem Identification variables to vary what the student has to do to identify the problem.
Explicitness (stated, partially identified, embedded, multiply-masked)	Describes the ambiguity of the problem to be identified. This variable determines the amount of work a student needs to do to identify the problem. Explicitness can assume four states: (a) stated—There is no ambiguity—the problem has been identified for the student; (b) embedded—There is no guidance as to which information sources would be helpful to identifying the problem. Therefore, the student needs to look at all information sources for problem identification. Information is presented at one time; (c) partially identified—Problem now has a bit of ambiguity, but problem statement still provides some guidance as to where to look to identify the problem; and (d) multiply-masked—Like an embedded problem, the student needs to look at all information sources for problem identification. However, the data are revealed as the student progresses through the scenario. Information is presented sequentially.
Barriers to getting information (none, allow for barriers)	Describes whether the task will make it hard to get information. *Barriers* can assume two states: (a) none—All of the information sources that are needed to identify the problem are made available to the student; and (b) allow for barriers—Information that is needed to solve the problem is difficult to access. The student will need to look at other sources to find out that information.
Time constraints (external, self-paced)	Time demands of the problem itself. *Time constraints* can assume two values: (a) external—Something in the problem space determines the time demand; and (b) self-paced—Completion of the task is not constrained by time.

Variable (and values)	Definitions
Consistency among information sources (consistent among information sources, allow for inconsistency among information sources)	Describes whether different information sources contradict each other. *Consistency <u>among</u> information sources* can assume two values: (a) consistent among information sources—The information from different sources is internally consistent.; and (b) allow for inconsistency among information sources—The information from one source contradicts information from another.
Consistency within information sources (consistent within information sources, allow inconsistency among information sources)	Describes whether information from one information source is changing over time (reflecting for example, faulty instrumentation, the degree of reliability, or information sources intentionally trying to mislead). *Consistency <u>within</u> information sources* can assume two values: (a) consistent within information sources—The information within a source is consistent; and (b) allow inconsistency among information sources—The information within a source is changing.
Accuracy of information sources (accurate information sources, allow inaccurate information sources)	Describes the degree to which information is correct. This typically requires an external referent. *Accuracy* can assume two values: (a) accurate information sources—The information within the information source is accurate; and (b) allow inaccurate information sources—The information within an information source is inaccurate.
Completeness of information sources (complete, allow for partial, allow for incomplete)	Describes the degree to which all information sources are provided to student—even those that are not useful for solving the problem. Note, this does not address whether the information is inaccurate. *Completeness* can assume three values: (a) complete—All of the information sources are made available; (b) allow for partial—Some of the information sources are made available. Information that is missing may be inferred from available sources; and (c) allow for incomplete—Information is missing and cannot be inferred from an available source.
Credibility of information sources (allow for not credible sources, low credibility, medium credibility, high credibility)	The trustworthiness of the information from source. *Credibility* can assume four values: (a) allow for not credible sources—Information from this source would not be credible and should be dismissed or ignored; (b) low credibility—Information from this source is highly subjective and would probably need additional information from another source to make use of it; (c) medium credibility—Information from this source, although somewhat subjective, is still fairly credible. For the most part, there is no reason to question information from this source; and (d) high credibility—There is never any reason to question information from this source—typically does not require interpretation or judgment.

Variable (and values)	Definitions
Relevancy of information sources (allow for no relevancy, low relevancy, medium relevancy, high relevancy of information source)	Whether or not an information source is relevant to identifying the problem. *Relevancy* can assume four values: (a) allow for no relevancy—Information is provided that is irrelevant to identifying the problem; (b) low relevancy—Information is only tangentially related to identifying the problem. Problem could be identified without this information; (c) medium relevancy—This information would be helpful with identifying the problem, but is not absolutely necessary. That is—identification of the problem does not absolutely require this information; and (d) high relevancy of information source—Without this information, identifying the problem would be impossible. Identification of the problem requires this information.
Number of information sources (zero, single, multiple)	The number of information sources that make up the problem space. *Number of information sources* can assume three values: (a) zero—No information sources available; (b) single—The problem information is contained in one source (e.g., a diagram); and (c) multiple—The problem information is contained in multiple sources (e.g., a diagram, graph, text, audio, etc.).
Prior knowledge (low prior knowledge, high prior knowledge)	The amount of domain knowledge required to identify the problem. This determines how much domain specific knowledge is required to solve the problem. *Prior knowledge* can assume two values: (a) low prior knowledge—Little knowledge or simply a basic grasp of the domain is required to solve the problem (factual knowledge); and (b) high prior knowledge—A good grasp of the domain is required to solve the problem (more conceptual knowledge).
PROBLEM CHARACTERISTICS	
Type of task (execute, fix, change usual sequence, improvise steps, combination of tasks)	Describes what type of problem-solving task is required of the student. *Type of task* can assume five values: (a) execute—Task requires the student to execute a known algorithm; (b) fix—Once a problem has been identified, the task is to find its solution; (c) change usual sequence—The task is to take a given solution to a problem and achieve the same goal using a different sequence of steps; (d) improvise steps—The task is to develop a novel solution to a problem; and (e) combination of tasks—The task may require a combination of three other types of tasks (*fix*, *change usual sequence*, and *improvise steps*).
SOLUTION STRATEGY	Describes what sort of strategies are required by the solution to the problem.
Steps (explicit course of action, non-specified course of action)	Whether the solution follows a prescribed sequence of actions. *Steps* can assume two values: (a) explicit course of action—There is a prescribed sequence of actions to solve the problem. Typically this is for tasks of type "fix"; and (b) non-specified course of action—There is no prescribed sequence of actions to solve the problem. Typically this is for tasks of type "improvise steps."

Variable (and values)	Definitions
Problem subdivision (required, not required)	Describes the degree to which you need to break apart the problem in order to solve it. *Problem subdivision* can assume two values: (a) required—Solving the problem requires breaking it up into smaller parts; and (b) not required—Solving the problem does not require that it be broken up into smaller parts.
Contingency planning (required, not required)	Describes whether the solution requires the student to have a backup plan. *Contingency* can assume two values: (a) required—Solution to the problem may not work because of a faulty assumption or incorrect hypothesis. Student will need to be able to recover from the error; and (b) not required—Once a solution is identified, it is guaranteed to work.
Help seeking (required, not required)	Describes whether the student needs to ask for help/assistance/support to solve the problem. *Help seeking* can assume two values: (a) required—Help seeking is necessary when, for example, information is withheld (e.g., permissions problem); the student does not have a strategy or does not know what to do with the information; it is not the student's domain or job; and (b) not required—If the student needs help to solve the problem, it is a due to a lack of knowledge rather than a function of the task.
Cognitive strategies (domain independent, domain dependent)	Describes whether or not a problem can be solved with a general algorithm or skill or requires specialized algorithms, or domain-specific methods and/or techniques. *Cognitive strategies* can assume two values: (a) domain independent—General algorithm independent of domain knowledge can be used to solve the problem (e.g., trial and error).; and (b) domain dependent—Domain-specific techniques.
SOLUTION CHARACTERISTICS	
Solution space (convergent, divergent)	Whether the problem has a known solution or an unknown solution. *Solution space* can assume two values: (a) convergent—Single right answer; and (b) divergent—Open-ended, with scoring criteria (typically judgmentally scored).
Solution correctness (multiple acceptable solutions, partially acceptable solution)	The degree to which something other than one solution is accepted. *Solution correctness* can assume two values: (a) multiple acceptable solutions—The problem has more than one correct answer; and (b) partially acceptable solution—Partially correct performance is used as a measure of competence.

Variable (and values)	Definitions
Sub-solution contingencies (sequential, non-sequential)	Whether the problem is sequential and conditional in nature. *Sub-solution contingencies* can assume two values: (a) sequential—Performance upon one aspect of the task determines what is presented in the next step of the task. The student will not know the result of his or her actions until it is applied. Information is dynamic; and (b) non-sequential—Performance does not determine a path the problem takes. The problem is solved in one step. Information is static.

Use-Case Scenarios

In the first part of the concept of operations we outline three scenarios typical of educational and training settings. These scenarios illustrate an instructor's view at the beginning, during, and at the end of a course with respect to gathering information about students. Following the scenarios we describe how our architecture supports these scenarios. We then provide two examples of the authoring process in a marksmanship context (one sound assessment design and one poorly designed assessment).

Scenario 1: Determining what students already know. The following scenario is typical of what instructors are confronted with at the beginning of instruction in general. This is a particular problem in training settings where instructors are confronted with waves of students with whom the instructors are unfamiliar and the trainees have varying backgrounds and experience.

> A new training course is starting and the students are coming with a wide range of backgrounds. The instructor wants to get a sense of what each student knows about the topic so that instruction can be adjusted to fit with the students' background. The instructor needs to decide on what to do given the constraints on instructional time, effort, and anticipated payoff for gathering information about the knowledge and skills of incoming students.

Scenario 2: Determining how much students are learning. The following scenario is typical of instruction in general. Two issues are central to this scenario: how instructors determine what students are learning and how timely that information is (i.e., can instructors act on the information). In practice, these issues are conditioned by the real-world constraints such as the amount of instructional time, the skill of the instructor in terms of instructional and assessment skill, and the capacity available (e.g., whether the instructor has available time, or an assistant to score tests).

> During the duration of the training course the instructor wants to get a sense of what each student is learning so that instruction can be adjusted to maximize student learning. In an ideal instructional situation, the instructor has information on how well each student is comprehending the content, how the student applies the knowledge to novel situations, what topics students find confusing, what topics are redundant, and what topics need more elaboration. Further, this information is accurate and at such a grain size that the instructor can take immediate action to adjust instruction and content. The instructor needs to decide on what to do given the constraints on instructional time,

effort, and anticipated payoff for gathering information about the knowledge and skills of current students.

Scenario 3: Determining the degree to which students have learned requisite skills and knowledge. An important question at the end of the course is whether students have acquired the knowledge and skills intended by instruction.

> Near the end of the training course the instructor wants to determine whether students have learned the content and skills specified in the course objectives. In an ideal instructional situation, the instructor has had ongoing information about the overall condition of the class and thus the instructor should have a reasonably accurate picture of whether the course objectives were met. Thus, any testing at the end of the course is verification that students have attained a particular level of skill and knowledge. Passing the test should reflect that the student has met the standard set for the course. The instructor needs to decide on whether to administer a test and if so, what students have to do to demonstrate content and skill mastery. Practical constraints include instructional time, effort, anticipated payoff, and whether the test is a good certification for the course.

In each of these scenarios, the typical set of options include the following (ordered by amount of in-class time):

1. Do nothing, assume students are learning
2. Whole-group questioning of students (show of hands)
3. Whole-group testing (short)
4. Whole-group testing (in depth)
5. Individualized questioning of each student
6. Individualized testing of each student

With the exception of the first three options, there is generally neither available class time to conduct the testing, nor the expertise to develop the test, nor the capacity to score the tests. Further, in many training and education situations, a highly desired outcome is a student who can use the knowledge and skills appropriately in novel and complex situations—problem solving and transfer. However, practical constraints and limitations in expertise suggest that testing students for problem-solving skill and transfer will not become common practice without a support structure that simultaneously makes the development and testing process more feasible (i.e., fit within the practical constraints of the classroom) and provides useful information to the instructor that can be acted on to improve instruction immediately.

Toward Authoring System Supports for Developing Problem-Solving Tasks

A basic assumption underlying the architecture is that the assessment structure be kept independent of the domain structure. The purpose of this separation is generalizability and scalability. With respect to generalizability, an independent assessment structure provides a substrate upon which all tasks are developed. A set of common design criteria increases the likelihood that the assessment tasks will inherit properties of the assessment structure. A set of desirable outcomes is that the system will allow the generation of tasks that are grounded in a theoretical framework, have a design traceable to a constraint network, and increase the overall quality of tasks by providing structure and guidance for authors with little knowledge of assessment design. The architecture described in this report reflects our current progress in developing such a system.

The key technical challenges for developing a domain-*independent* representation for the assessment of problem solving are:

- Identifying the key variables that represent the domain of problem solving with respect to assessment.
- Identifying the set of states the variables can assume.
- Codifying the relationships among variables.

The key technical challenges for developing a domain-*dependent* representation are:

- Instantiating, in a particular domain, the domain-dependent correlates of the assessment variables.
- Ensuring adequate domain coverage to provide content and context rich enough to exercise examinees' problem-solving skills.

An essential component underlying the system architecture is a constraint network. In a constraint network, nodes are variables that can assume a range of values and the topology specifies how the variables and values are related (Montanari, 1974). To support assessment design, the constraint network will codify the major concepts and relations that underlie high-quality assessments (e.g., Linn et al., 1991). A constraint-processing engine will evaluate the fit between the user's design and the assessment ontology in the context of the domain ontology, and alert the user of constraint violations as well as options in the design that would satisfy all constraints (Montanari, 1974).

Assessment Design as Constraint Satisfaction

For the purpose of assessment design, a constraint network can be used to explicitly represent an assessment model and provide a description of assessment parameters, the constraints governing relationships among the parameters, and computational access to the parameters and constraints. This representation can be used, for example, to provide guidance to assessment authors as they design assessments for particular purposes under particular constraints. That is, the explicit structure is of very high utility because it allows the enforcement of a common and consistent framework. This structure can be leveraged to assist assessment authors (particularly non-experts) in designing assessments. Assessment authoring support could be in the form of (a) aiding assessment authors to populate the assessment ontology with values specific to the users' purposes, and (b) system constraint checking that would ensure that assessment authors are alerted to incompatible values.

The specification of the different components of the assessment is the critical first step. We have defined the top-level components and allowable states as shown in Table 1. Table 2 shows 18 constraints that capture the model described in Baker and O'Neil (2002). These constraints specify the allowable conditions of problem-solving task design. For each constraint, the variable and value(s) are listed in the following format: *<category>::<category variable>::(possible values)*. A constraint is satisfied if all variables are set to the specified values. A design is acceptable if all constraints are satisfied. Note that the model in Table 2 represents one point of view of a problem-solving assessment model. Other perspectives will have different representations and constraints. While this idea is not new (e.g., see Baker, 1997), what has changed is the availability of computational tools that make feasible the capturing and processing of the model computationally.

Table 2

Constraints Associated With Designing a Problem-Solving Task

Set No.	Constraints
1	COGNITIVE DEMAND::-::(reasoning) PROBLEM IDENTIFICATION::explicitness::(partially identified, embedded, multiply-masked) PROBLEM IDENTIFICATION::barriers to getting information::(allow for barriers) PROBLEM IDENTIFICATION::prior knowledge::(low prior knowledge, high prior knowledge)
2	COGNITIVE DEMAND::-::(content understanding) PROBLEM IDENTIFICATION::explicitness::(partially identified, embedded, multiply-masked) PROBLEM IDENTIFICATION::barriers to getting information::(none) PROBLEM IDENTIFICATION::prior knowledge::(high prior knowledge)
3	COGNITIVE DEMAND::-::(factual knowledge) PROBLEM IDENTIFICATION::explicitness::(stated) PROBLEM IDENTIFICATION::prior knowledge::(high prior knowledge)
4	PROBLEM IDENTIFICATION::explicitness::(stated, embedded) PROBLEM IDENTIFICATION::prior knowledge::(low prior knowledge)
5	PROBLEM IDENTIFICATION::explicitness::(stated, embedded) SOLUTION CHARACTERISTICS::sub-solution contingencies::(non-sequential)
6	PROBLEM CHARACTERIZATION::type of task::(improvise steps, combination of tasks) SOLUTION STRATEGY::steps::(non-specified course of action)
7	SOLUTION STRATEGY::contingency planning::(not required) SOLUTION CHARACTERISTICS::solution space::(convergent) SOLUTION CHARACTERISTICS::sub-solution contingencies::(non-sequential)
8	SOLUTION STRATEGY::contingency planning::(required) SOLUTION CHARACTERISTICS::sub-solution contingencies::(sequential)
9	PROBLEM CHARACTERIZATION::type of task::(execute, fix) SOLUTION STRATEGY::steps::(explicit course of action)
10	PROBLEM CHARACTERIZATION::type of task::(execute, fix) SOLUTION CHARACTERISTICS::solution space::(convergent) SOLUTION CHARACTERISTICS::solution correctness::(single acceptable solution)
11	PROBLEM IDENTIFICATION::prior knowledge::(low prior knowledge) SOLUTION STRATEGY::cognitive strategies::(domain independent)
12	PROBLEM IDENTIFICATION::prior knowledge::(high prior knowledge)

Set No.	Constraints
	PROBLEM CHARACTERIZATION::type of task::(execute) SOLUTION STRATEGY::cognitive strategies::(domain dependent)
13	PROBLEM IDENTIFICATION::time constraints::(self-paced) SOLUTION STRATEGY::cognitive strategies::(domain dependent)
14	PROBLEM IDENTIFICATION::time constraints::(external) SOLUTION STRATEGY::cognitive strategies::(domain independent)
15	PROBLEM IDENTIFICATION::explicitness::(stated) PROBLEM IDENTIFICATION::number of information sources::(zero)
16	PROBLEM IDENTIFICATION::explicitness::(partially identified, embedded, multiply-masked) PROBLEM IDENTIFICATION::number of information sources::(single, multiple)
17	PROBLEM IDENTIFICATION::explicitness::(stated, partially identified) SOLUTION STRATEGY::problem subdivision::(not required)
18	PROBLEM IDENTIFICATION::explicitness::(embedded, multiply-masked) SOLUTION STRATEGY::problem subdivision::(required)

Applying the Framework: Examples From Two Domains

In an ideal assessment design, the specifications of the task would be determined by how the information from the assessment would be used. The purpose of the assessment should drive the complexity of the problem as well as define what sort of strategies are necessary to solve the problem. The following examples demonstrate this notion of matching the task specifications with the assessment purpose (the appendix contains an additional example). Using the scenarios mentioned earlier in the context of a marksmanship coach course which trains Marines on how to be marksmanship coaches, we specify the assessment component values with respect to the goal of the assessment.

Example 1. Determining What Students Already Know

This first example demonstrates what the authoring process of a good assessment would look like for Scenario 1, "determining what students already know." This is from the view of an instructor administering the assessment at the beginning of a course trying to get a sense of how much knowledge the students have about the marksmanship coaching process. The task specifications of this example are found in Table 3.

> A new course is starting and the students are coming with a wide range of backgrounds. The instructor wants to get a general measure of a student's knowledge about marksmanship and the marksmanship coaching process. The assessment is administered prior to any instruction in order to make adjustments to the curriculum. The task administered to the student involves identifying a shooter's problem by looking at various information sources (e.g., data book, target, shooter's position, weapon settings, etc.) and prescribing its appropriate solution under the same time constraints as a coach on a range (approximately 5 minutes). In this particular example, the shooter's problem is with windage settings—the shooter had not properly compensated for wind prior to firing. Therefore, all of the shots are forming a group to the left of the center of the target.
>
> **Critique.** Very little guidance on identifying or solving the shooter's problem will be provided to the student. Therefore, performance on this task is highly dependent upon the student's prior knowledge, with students who know more about marksmanship expected to do better than those who do not know as much. Because the instructor is not particularly interested in whether a student knows what to do if there is difficulty with an information source (either because of availability, accuracy, completeness, or consistency) all of the sources are available (including those that may be irrelevant to identifying and solving the problem), accurate, and consistent both within and among each other. Once the shooter's problem has been identified, the assessment requires that the student "fix" the problem (adjust the shooter's windage). The remedy is fairly straightforward, so the student is not required to come up with a creative solution.

Example 2. Determining the Degree to Which Students Have Learned Requisite Skills and Knowledge

This example is what the authoring process of a *poorly* designed assessment would look like for Scenario 3, "determining the degree to which the students have learned requisite skills and knowledge." In this case, the instructor administers the assessment task at the end of the course to see whether the students are prepared to begin coaching on the range. The task specifications of this example are found in Table 3. The aspects of the task that contribute to its poor design are italicized.

> The coach course has been completed and the instructor is unsure if several of the students are prepared to coach on the range. The assessment's purpose is to see in which areas the students are still having difficulty. The instructor is particularly concerned that students do not know what to do if a piece of information is missing in order to make a diagnosis.

Critique. This design has two key flaws: (a) explicitness of problem identification and (b) the quality of information sources. Problem identification has very little ambiguity which does not allow adequate demonstration of knowledge. The task guides the student as to where they should look to figure out what is the shooter's problem (instructor chooses a problem statement that says, *"Shooter's problem has something to do with her weapon"*). The instructor is also looking for the ability to deal with complications such as missing information because coaches often encounter this situation when it is raining. However, the instructor designed a task that had *no barriers to getting information,* and *completeness of information sources.* By choosing these states, the assessment will make everything available to the student and the instructor will not be able to measure competence in dealing with missing information.

Table 3

Task Specifications of a Diagnostic Assessment for Coaching Rifle Marksmanship

Variable	Example	Value	Comments
ASSESSMENT PURPOSE	1	diagnostic	Assessment is given at the beginning of the coach course.
	2	diagnostic	Assessment is given at the end of the coach course.
COGNITIVE DEMAND	1	content understanding	Problem requires the student to use content knowledge.
	2	reasoning	Student is asked to use multiple information sources to identify a problem. Problem requires the student use both content knowledge and reasoning skills.
PROBLEM IDENTIFICATION Explicitness	1	embedded	Student is told that the shooter is having difficulty hitting the center of the target.
	2	partially identified	Student is told that the shooter's problem is with the weapon. *The instructor wants to know what difficulties a student is still having. Partially identifying the problem provides too much guidance.*
Barriers to getting info.	1	none	Student can access all information sources.
	2	none	Student can access all information sources. *Allowing access to all pieces of information will not adequately measure a student's ability to deal with missing information.*
Time constraints	1, 2	external	Student is given five minutes to complete the assessment—mimics the time a coach is given to identify a shooter's problem and fix it.
Consistency among info. sources	1, 2	consistent	For example, the information in the data book is consistent with the information from the target.
Consistency within info. sources	1, 2	consistent	For example, the wind values are consistent. From the time before firing the shots commence to the end of the stage, the wind values remain the same.
Accuracy of info. sources	1, 2	accurate	All of the information is accurate.
Completeness of info. sources	1	complete	Everything is presented in the scenario, including less useful information.

Variable	Example	Value	Comments
	2	complete	All of the sources are complete and presented to the student. *Having everything complete and presented to the student will not allow an instructor to measure a student's ability to deal with missing information.*
Credibility of info. sources	1, 2	allow for not credible source	Anything a shooter has to maintain a record of requires a shooter's interpretation, thus this information is suspect.
Relevancy of info. sources	1, 2	allow for no relevancy	A shooter's ethnicity is presented to the student. This information will not help a student identify a shooter's problem.
No. of info. sources	1, 2	multiple	In this domain, assessments will always have multiple information sources.
Prior knowledge	1	—	Assessment is taken at the beginning of the course so it is unknown how much students know about the coaching process.
	2	high	Assessment is taken at the end of the course so it is assumed that the student have acquired the requisite knowledge about the coaching process.
PROBLEM CHARACTERISTICS Type of task	1, 2	fix	Once the student identifies the shooter's problem, the task requires that the student fix the windage settings.
SOLUTION STRATEGY Steps	1, 2	explicit	The solution to the problem requires fixing a shooter's windage settings by adding clicks to the right.
Problem subdivision	1, 2	N/A	No subdivision is required to solve the problem.
Contingency planning	1	not required	Does not apply—student knows that by adjusting the settings to compensate for the wind will cause the shots to hit the center of the target.
	2	required	Student will need to have a backup strategy in the event that the applied solution to compensate for a missing data book fails to solve the problem.
Help seeking	1, 2	not required	The student does not need to seek help to diagnose and fix a shooter's problem.
Cognitive strategies	1, 2	domain dependent	Problem needs to be solved using specific marksmanship coaching strategies.
SOLUTION CHARACTERISTICS Solution space	1	convergent	The problem has a known solution: to adjust the windage settings.

Variable	Example	Value	Comments
	2	divergent	The problem does not have a known solution.
Solution correctness	1, 2	N/A	
Sub-solution contingencies	1, 2	non-sequential	Progression through the problem is not sequential.

Example 3. Mismatch Between Assessment Purpose and Assessment Design

This example comes from an assessment used in a mathematics lesson (reported in Koency, 2000). This example illustrates how the intended purpose of the assessment can be undermined by the interaction between prompt and information source. In this case, the instructor administered the task at the end of the unit to gather information on whether students had mastered fractions. The teacher's objectives of the unit were: (a) students will understand percents as an alternative way of representing fractions with a denominator of 100 and (b) students will build an understanding of the relationships between fractions, decimals, and percents. The use of the assessment and critique are described next.

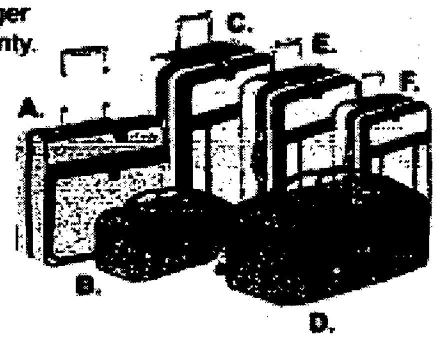
> At the end of the unit, the teacher gave students the worksheet shown in Figure 1. The students were given the prompt "Which one is the better sale (A or B)?" and were asked to provide a written explanation justifying their choice. Students were required to record the sale price for each piece of luggage under the "Sale" column. Using the variables in Table 1, the task can be described as shown in Table 4. The major aspects of the task are: (a) assessment purpose (*certification*), (b) cognitive demand (*content understanding*), (c) explicitness of problem identification (*embedded*), and (d) relevancy of information sources (*allow for no relevancy of information sources*).
>
> **Critique.** The teacher's purpose was to assess whether the students understood the meaning of percentages (*certification* and *content understanding*) by using percentage arithmetic to calculate the savings and make a quantity comparison between two values. Students were asked to report the various strategies they employed to find the new luggage prices. However, the teacher designed a task that had *embedded* problem identification that contained information sources that *allowed for no relevancy* types. The prompt in conjunction with the information on the worksheet created a potentially ambiguous situation. The problem asked students to determine "which one is the better sale" but the assessment scenario contained a number of irrelevant information sources (e.g., warranty information, luggage features, wheels) that were not needed to solve the problem. An *embedded* problem is appropriate for tasks requiring the cognitive demand of *reasoning*.
>
> The ambiguity lay with the potential for a student to interpret the task as one of considering the additional information (e.g., warranty and luggage features) with cost—a very reasonable approach, especially given the "authentic" context of the task. Poor performance on the task may be due to a misinterpreting what the problem is asking for rather than not knowing the material. This source of construct-irrelevant variance subverts the intended purpose of the assessment, to provide information on whether students have attained the objectives.

Luggage Sale!

A. Our Most Exclusive Brand—25% off

Tightly woven softside luggage in bright red or green. Larger pieces on wheels for easy maneuverability. 10-year warranty.

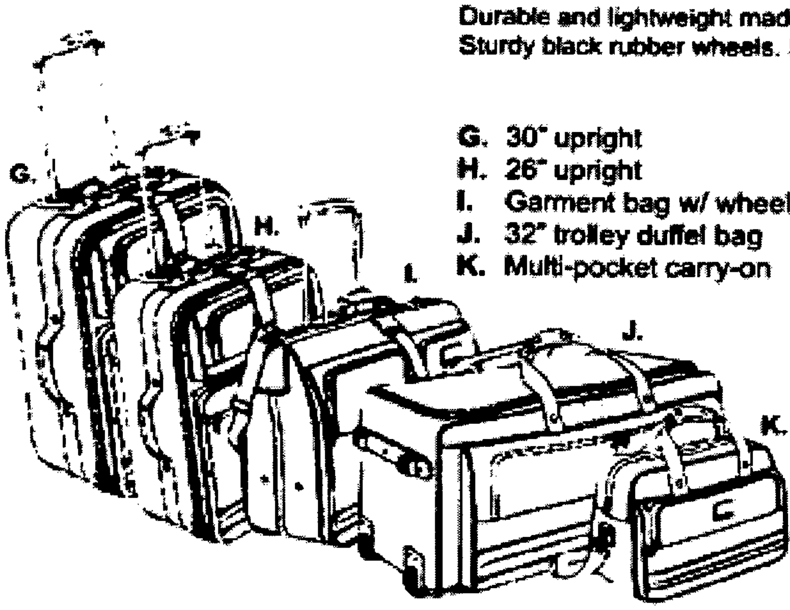
	Reg.	Sale
A. Wheeled garment bag	$228	
B. Attachable multi-pocket tote	$54	
C. 29" upright pullman	$215	
D. Multi-pocket carry-on	$80	
E. 26" upright pullman	$215	
F. 22" upright suiter	$175	



B. Famous Maker Luggage—50% off

Durable and lightweight made of 1200-denier polyester. Sturdy black rubber wheels. 5-year warranty.

	Reg.	Sale
G. 30" upright	$260	
H. 26" upright	$200	
I. Garment bag w/ wheels	$260	
J. 32" trolley duffel bag	$200	
K. Multi-pocket carry-on	$70	



Which one is the better sale? (A or B). Write your explanation on the back of this page.

Figure 1. Example 3 information source.

Example 4. Correcting the Mismatch Between Teacher's Assessment Purpose and Assessment Design in Example 3

The problem in Example 3 lies in the four incompatible values of the assessment design: certification (*purpose*), cognitive demand (*content understanding*), explicitness (*embedded*), and relevancy of information sources (*allow for no relevancy of information source*). To fix the task design in Example 3, the explicitness of the problem to be identified should be stated. An example of an explicit prompt is: "Based on the calculated savings, determine which offer is the better sale." The prompt makes explicit that students should base their decision only on the calculated savings. Presumably, students will apply only their knowledge of percentages (*content understanding*) to carry out the task, rather than considering the extraneous information contained in the scenario (*reasoning*). The slight adjustment to the prompt aligns the assessment purpose with cognitive demand (*content understanding*) and problem identification explicitness (*stated*). The revised task specifications are shown as Example 4 in Table 4.

Table 4

Task Specifications of a Diagnostic Assessment for Mathematics

Variable	Example	Value	Comments
ASSESSMENT PURPOSE	3, 4	certification	Assessment is given at the end of the lesson. Teacher wants to assess mastery of using percentages.
COGNITIVE DEMAND	3	content understanding	Problem requires the student to use content knowledge. *Incompatible with the vague problem statement prompt (embedded).*
	4	content understanding	Problem requires the student to use content knowledge.
PROBLEM IDENTIFICATION Explicitness	3	embedded	Student asked to identify which is the better sale in the scenario contained in Figure 1. *Problem statement prompt asked the student to identify the better sale. Given the information sources contained in the scenario, a student may use the irrelevant information (e.g., warranty details, luggage features) to determine which is the better sale.*
	4	stated	Problem statement prompt asks the student to calculate the savings and use the results to identify which is the better sale in the scenario contained in Figure 1.
Barriers to getting info.	3, 4	none	Students can access all information sources.
Time constraints	3, 4	self-paced	Students are not given a time constraint other than the class period.
Consistency among info. sources	3, 4	consistent	Information from sources are not dependent upon each other.
Consistency within info. sources	3, 4	consistent	There are no changing values within an information source.
Accuracy of info. sources	3, 4	accurate	All of the information is accurate.
Completeness of info. sources	3, 4	complete	Everything is presented in the scenario, including less useful information.
Credibility of info. sources	3, 4	allow for high credibility	All of the information in the scenario is credible.
Relevancy of info. sources	3	allow for no relevancy	Information such as the warranty information and luggage descriptions (e.g., wheels) are included which are not necessary to solve the problem. *Having irrelevant information that might cause confusion about how to solve the problem may cause a mismatch between a student's performance and the student's actual knowledge.*

Variable	Example	Value	Comments
	4	allow for no relevancy	Information such as the warranty information and luggage descriptions (e.g., wheels) are included which are not necessary to solve the problem. *Because the problem statement is explicit, allowing for irrelevant information should not confusion.*
No. of info. sources	3, 4	multiple	Prices, percentages off, warranty information, luggage descriptions, etc.
Prior knowledge	3, 4	high	Assessment is taken the end of the unit so it is assumed that the student fully understands percentages.
PROBLEM CHARACTERISTICS Type of task	3, 4	execute	Problem requires that the students execute known algorithms to solve the problem.
SOLUTION STRATEGY Steps	3, 4	explicit course of action	There is a known solution to the problem.
Problem subdivision	3, 4	required	Students must first calculate the savings. Then, given the results, students must make a comparison to determine which sale (20% off or 50% off) is the better sale.
Contingency planning	3, 4	not required	Does not apply.
Help seeking	3, 4	not required	Does not apply.
Cognitive strategies	3, 4	domain dependent	Use knowledge of percentages to solve the problem. Cannot be solved using weak methods such as trial and error or means-ends analysis.
SOLUTION CHARACTERISTICS Solution space	3, 4	convergent	The problem has a known solution: to calculate the savings using percentages.
Solution correctness	3, 4	multiple acceptable solutions	Does not apply.
Sub-solution contingencies	3, 4	non-sequential	Data in the scenario are not presented sequentially.

Discussion and Next Steps

In this report we outlined an approach to the design of an authoring system for problem-solving tasks. A key feature to support evaluation of assessment designs is the use of a constraint network to capture the allowable relations among the assessment model variables. The basic idea is that major assessment and task variables can assume a fixed set of values, and that constraints among the variables define allowable relations or conditions. A task design can then be validated by checking for constraint violations.

Technologies that can support the assessment authoring design process seem particularly promising because of the nature of the anticipated users: non-experts who lack breadth and depth of knowledge of assessment. A constraint-based authoring system can impose structure on the authoring process, focusing users' attention on the important variables underlying the assessment task, as well as verifying that user-specified values are consistent with the underlying assessment model. The advantage of a constraint-based approach is a tighter coupling between the assessment model and the instantiated task. Presumably, the more the task design adheres to the assessment model, the higher the task quality—this would make the authoring system particularly useful for novice assessment authors.

While we are confident that the technology component (i.e., constraint processing) exists, the larger assessment issue is whether the domain can be captured in terms of variables and states, and whether such technology-enabled solutions result in higher quality assessments (Baker, 2003). We have attempted to illustrate how problem-solving tasks could be represented with constraints and how constraints could "catch" design flaws, but more work is needed to test the notion of constraints applied to assessment design, the variety of constraints, their interactions, and the implications for designing tasks.

Next steps for this work include gathering evidence on the degree to which our framework yields judgments similar to those of experts for good and poor assessments. Constraint violations should be detected for poor tasks and be absent for good assessments. Such evidence would support the interpretation that the framework was capturing meaningful aspects of the assessment design. A second test is to examine how the authoring system would work for authors of differing backgrounds. For example, one set of comparisons is between novice test designers with and without the authoring framework (e.g., typical classroom instructors). Effectiveness of our system would be evidenced by higher quality assessments created by novices using our system compared to those not using our system.

As assessment design moves from craft knowledge to an engineering discipline, models and tools will be needed to facilitate the systematic development of assessments. We have presented one approach that could be implemented using constraint networks to support the assessment design process.

References

Baker, E. L. (1997). Model-based performance assessment. *Theory Into Practice, 36,* 247-254.

Baker, E. L. (2001). Technology: How do we know it works? In J. Willis (Series Ed.) & W. F. Heinecke & L. Blasi (Vol. Eds.), *Research methods for educational technology: Vol. 1. Methods of evaluating educational technology* (pp. 77-84). Greenwich, CT: Information Age.

Baker, E. L. (2002a). Design of automated authoring systems for tests. In National Research Council, Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education (Eds.), *Technology and assessment: Thinking ahead: Proceedings from a workshop* (pp. 79-89). Washington, DC: National Academy Press.

Baker, E. L. (2002b). Teacher use of formal assessment in the classroom. In W. Hawley (Ed.) (with D. L. Rollie), *The keys to effective schools: Educational reform as continuous improvement* (pp. 56-64). Thousand Oaks, CA: Corwin.

Baker, E. L. (2003). Reflections on technology-enhanced assessment. *Assessment in Education, 10,* 421–425.

Baker, E. L., Gearhart, M., & Herman, J. L. (1994). Evaluating the Apple Classrooms of Tomorrow[SM]. In E. L. Baker and H. F. O'Neil, Jr. (Eds.), *Technology assessment in education and training* (Vol. 1, pp. 173-197). Hillsdale, NJ: Lawrence Erlbaum.

Baker, E. L., & Herman, J. L. (2000, February). *Technology and evaluation.* Paper prepared for the presentation "New models of technology sensitive evaluation: Giving up old program evaluation ideas" at the Technology Design Meeting, SRI International, Menlo Park, CA.

Baker, E. L., Herman, J. L., & Gearhart, M. (1996). Does technology work in schools? Why evaluation cannot tell the full story. In C. Fisher, D. C. Dwyer, & K. Yocam (Eds.), *Education and technology: Reflections on computing in classrooms* (pp. 185-202). San Francisco: Jossey-Bass.

Baker, E. L., & O'Neil, H. F., Jr. (2002). Measuring problem solving in computer environments: current and future states. *Computers in Human Behavior, 18,* 609–622.

Baker, E. L., & O'Neil, H. F., Jr. (2003). Evaluation and research for technology: not just playing around. *Evaluation and Program Planning, 26,* 169–176.

CAPA. (2000). *Computer-assisted personalized approach (CAPA)* [On-line]. Available: http://capa4.lite.msu.edu/homepage/#docs

Chen, E. J., Chung, G. K. W. K, Klein, D. C. D., de Vries, L. F., & Burnam, B. (2000). *Year 2 Evaluation Findings* (Final deliverable to donor [anonymous]). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, G. K. W. K., Baker, E. L., & Cheak, A. M. (2001). *Knowledge mapper authoring system prototype.* (Final deliverable to OERI). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, G. K. W. K., Klein, D. C. D., & Baker, E. L. (2000). Year 1 evaluation report of the Advanced Educational Technologies Program (AETP) (Final Deliverable to donor [anonymous]. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, G. K. W. K., Klein, D. C. D., Herl, H. E., & Bewley, W. (2001). *System specifications for the design of a knowledge map authoring system* (Draft deliverable to OERI). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Clark, V. (2003, January 3). *CNO Guidance for 2003.* Retrieved from http://www.chinfo.navy.mil/navpalib/cno/clark-guidance2003.html.

Educational Testing Service. (2000). Assessment Wizard: Software for designing and sharing student assessments [Computer software]. Princeton, NJ: Author.

Eisenhower National Clearinghouse. (2001). *The Eisenhower national clearinghouse for mathematics and science education* [On-line]. Available: http://www.enc.org/topics/assessment/selections/

Gearhart, M., Herman, J. L., Baker, E. L., Novak, J. R., & Whittaker, A. K. (1994). A new mirror for the classroom: A technology-based tool for documenting the impact of technology on instruction. In E. L. Baker and H. F. O'Neil, Jr. (Eds.), *Technology assessment in education and training* (Vol. 1, pp. 153-172). Hillsdale, NJ: Lawrence Erlbaum.

Harvard Graduate School of Education. (2001). *Balanced assessment in mathematics* [On-line]. Available: http://balancedassessment.gse.harvard.edu/

Koency, G. (2000). *Reliability and validity of rational number problem-solving assessments.* Unpublished doctoral dissertation, University of California, Los Angeles.

Learning Manager. (2005). *The learning manager* [On-line]. Available: http://www.thelearningmanager.com/

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50,* 741–749.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., & Almond, R. G. (1999). A cognitive task analysis with implications for designing simulation-based performance assessment. *Computers in Human Behavior, 15,* 335-374.

Montanari, U. (1974). Networks of constraints: Fundamental properties and applications to picture processing. *Information Sciences, 7,* 95–132.

NPDC Reusable Content Development Industry Days. (March 2003). Posted at http://www.npdc.navy.mil.

O'Neil, H. F., Jr., & Baker, E. L. (1993). Evaluation models for expert systems in military training. In R. J. Seidel & P. R. Chatelier (Eds.), *Advanced technologies applied to training design* (pp. 299-310). New York: Plenum Press.

O'Neil, H. F., Jr., & Baker, E. L. (1997). A technology-based authoring system for assessment. In S. Dijkstra, N. M. Seel, F. Schott, & R. D. Tennyson (Eds.), *Instructional design: International perspectives* (pp. 113–133). Mahwah, NJ: Erlbaum.

Osmundson, E., Jeffries, C., & Herman, J. (1998, April). *DEEMS: A Web-based Tool for science and math program evaluations*. Demonstration presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Queensland Association of Mathematics Teachers. (1998). *Secondary mathematics assessment and resource database* [On-line]. Available: http://smard.cqu.edu.au/

Quellmalz, E., Hinojosa, T., Hinojosa, L., & Schank, P. (2000). *Performance assessment links in science (PALS): An on-line resource library* (Final report to NSF). Menlo Park, CA: SRI, Center for Technology in Learning.

QuestionMark. (2001). *QuestionMark* [On-line]. Available: http://www.questionmark.com/us/home.htm

TRIADS. (2001). *Tripartite interactive assessment delivery system (TRIADS)* [On-line]. Available: http://www.derby.ac.uk/assess/newdemo/mainmenu.html van der Linden, W. J. (2005). *Linear models for optimal test design.* New York: Springer.

WebCT. (2001). *WebCT* [On-line]. Available: http://www.webct.com/

Appendix

Worked Examples

Variable and Value	Marksmanship Coaching Example	Parenting Class (Sick Baby) Example
PURPOSE		
Diagnostic	Assessment is given at the beginning of coach's course. In this example, an instructor administers the assessment to get a baseline of the student's knowledge of the coaching process. If the students have difficulty using the data book as an information source, instruction might focus on that topic.	Assessment is given at the beginning of parenting class. In this example, an instructor administers the assessment to get a baseline of the parent's knowledge of what do to when faced with a sick baby. Performance on this assessment will help the instructor guide instruction.
Certification	Assessment is taken at end of coach's course. The assessment is administered simply to see if the student mastered the coach's course. Performance is not used to modify instruction.	Assessment is taken at end of parenting class. The assessment is administered simply to see if the student learned what to do when faced with a sick baby. Performance is not used to modify instruction.
COGNITIVE DEMAND		
Self-regulation	Student is told that the shooter is having difficulty with hitting center. In this example, a student will be navigating through the problem space sequentially. The next state of the problem is dependent upon what the student does. Therefore, the student must constantly monitor what he does throughout the process and reflect upon past actions.	Student is told that the baby is sick. In this example, a student will be navigating through the problem space sequentially. The next state of the problem is dependent upon what the student does. Therefore, the student must constantly monitor what he does throughout the process and reflect upon past actions.
Reasoning	Student is told that the shooter is having difficulty with hitting center. The task requires examining all the information sources to identify the problem.	Student is told that the baby is sick. The task requires examining all the information sources to identify the problem.
Content understanding	Student is told that the shooter's problem may be with the shooter's positioning. The student needs to demonstrate understanding of shooting positions to solve the problem.	Student is given the baby's symptoms (fever, change in eating habits, diarrhea, and fussiness). The student needs to demonstrate understanding of how to make a diagnosis given a list of symptoms.

Variable and Value	Marksmanship Coaching Example	Parenting Class (Sick Baby) Example
Factual knowledge	Student is told that the shooter has incorrect windage settings. The student needs to demonstrate knowledge of how to fix incorrect windage settings (executing a known algorithm).	Student is told that the baby has the flu. The student needs to demonstrate knowledge of how to resolve the problem—call the doctor and alleviate flu symptoms (bring down fever, keep baby hydrated, etc.).
PROBLEM IDENTIFICATION		
Explicitness		
Stated	Student is told that the shooter has incorrect windage settings. Student is told what the shooter's problem is (no problem identification required). Task requires simply resolving the problem (make the appropriate rear sight windage settings adjustment).	Student is told that the baby has the flu. Student is told what is wrong with the baby (no problem identification required). Task requires simply resolving the problem—call the doctor and alleviate flu symptoms (bring down fever, keep baby hydrated, etc.).
Partially identified	Student is told that the shooter's problem may be with his weapon. The student only needs to look at information sources that are relevant to identifying the source of the weapon problem to figure out what needs to be fixed. For example, these information sources are the data book, target, and the weapon's sight settings. Information from other information sources will not be helpful with identifying the problem.	Student is given the baby's symptoms (fever, change in eating habits, diarrhea, and fussiness). In this example, problem identification entails taking the given symptoms and making a diagnosis. Student does not need to investigate the problem space for symptoms.

Variable and Value	Marksmanship Coaching Example	Parenting Class (Sick Baby) Example
Embedded	Student is told that the shooter is having difficulty with hitting center. The task requires examining all the information sources to identify the problem. In this example, all the information is internally consistent with the diagnosis of a windage settings problem. Aside from the initial shot, all of the shots are grouped at the 3 o'clock border. The range flags indicate no breeze coming from the left. Examination of the Before Firing section of the data book reveals that there was an existing breeze from the left but that the shooter did not adjust the sights for the wind.	Student is told that the baby is sick. In this example, all the information is internally consistent with the diagnosis of the flu. Baby has diarrhea and vomiting, and the symptoms are occurring during high flu season.
Multiply-masked	Student is told that the shooter is having difficulty hitting center. In this example, a student will be navigating through the problem space sequentially. The student will see the information sources before the shot is fired, as the shot is being fired, and after the shot is fired.	Student is told that the baby is sick. In this example, a student will be navigating through the problem space (temperature, eating habits, bowel movements, etc.) sequentially.
Barriers to getting information		
None	No barriers to the information. The student can view all of the information sources.	No barriers to the information. The student can view all of the information sources.
Allow for barriers	Bullet strikes are obscured on the target. Because the bullet strikes are obscured on the target, coach will need to examine the data book to see where the shots hit.	Baby cannot talk or squirming so much that temperature cannot be taken. Baby's inability to verbally communicates makes it difficult to grasp what pain he is in—the student needs to rely on nonverbal cues—crying, kicking, etc. The temperature is important—if baby's movement makes it impossible to take, student may have to determine whether he has a fever by touching his forehead.

Variable and Value	Marksmanship Coaching Example	Parenting Class (Sick Baby) Example
Time constraints		
External	Five minutes to complete the assessment. The assessment is designed to mimic the time a coach is given to identify a shooter's problem and to correct it.	Student is asked to make the assessment in less than ten minutes. Assessment is designed to mimic the stress that a crying baby will cause to a parent.
Self-paced	Does not apply. Assessments will always be constrained by time.	Does not apply. Diagnosis of a sick baby is always constrained by time.
Consistency among information sources		
Consistent	Information in the data book is consistent with the information from the target. For example, the shots plotted in the data book match up with the shots on the target.	Baby feels warm and has a temperature of 101.3°. The baby's temperature and body heat can both be indicators of whether a fever is present, and information from both sources is consistent.
Allow inconsistencies	Information in the data book is inconsistent with the information from the target. For example, the data book shows that the shot hit the center of the target but the actual target strike hits high and left.	Baby feels warm and has a temperature of 98.7°. The baby's temperature and body heat can both be indicators of whether a fever is present, and information from both sources is inconsistent.
Consistency within information sources		
Consistent	Wind values are consistent. From the time before the shots are fired to the end of the stage, the wind values remain the same.	Temperature readings are consistent. The baby's temperature is always 101°.
Allow inconsistencies	Wind values are inconsistent. The weather is changing over the course of fire.	Temperature readings are inconsistent. Sometimes the temperature is 101° and other times it is 98.7°.

Variable and Value	Marksmanship Coaching Example	Parenting Class (Sick Baby) Example
Accuracy of information sources		
Accurate	All of the information from the information sources is accurate. The information that might be susceptible to inaccuracies (e.g., data book, shooter's self-reported information, etc.) is accurate.	All of the information from the information sources is accurate.
Allow inaccuracies	Does not apply. If it's inaccurate, then it's inconsistent. There is no external referent to judge accuracy.	The thermometer is broken. The information from the thermometer is inaccurate.
Completeness of information sources		
Complete	Everything is presented in the scenario, including less useful information such as marital status, age, etc. All the information in the domain of marksmanship is made available to the student.	Everything is presented in the scenario, including less useful information such as eye color, hair color, what the baby is wearing, etc. All the information about the baby and his surroundings is made available to the student.
Allow for partial	The data book is made available but the target is not. The student has to infer from the data book where the shots hit the target because the target is not available.	You can feel that the baby is extremely warm (enough to suggest a fever) but your thermometer is broken. The parent has to infer that the baby has a fever because the thermometer is not available.
Allow for incomplete	Shooter is describing the situation to the coach after shooting. All of the information sources are missing—there is no data book, target, range, or weapon.	The baby has a sore throat but cannot communicate the pain because of the inability to talk. The parent cannot tell that the baby has a sore throat as well because the baby cannot talk. This is an essential piece of information that cannot be inferred from an available source.

Variable and Value	Marksmanship Coaching Example	Parenting Class (Sick Baby) Example
Credibility of information sources		
Allow for not credible sources	Anything a shooter has to maintain a record of. Because this information requires a shooter's interpretation, the information is suspect depending on the shooter's past performance and prior knowledge.	Your mom tells you your baby must be sick because she woke up with a bad feeling. A "whim" from an overbearing grandparent should be dismissed or ignored.
Allow for low credibility		Your wife tells you that the baby feels warm. Because this is not the most reliable way to take a baby's temperature, this information is of low credibility. To ensure its accuracy, one must take the baby's temperature.
Allow for medium credibility		Baby is fussy. Typically parents know when their baby is more fussy than usual, but this information still requires judgment.
Allow for high credibility		Symptoms such as coughing. Noting a baby coughing does not require interpretation or judgment.
Relevancy of information sources		
Allow for no relevancy of information source	A shooter's ethnicity. Information about ethnicity would not help a student identify a shooter's problem.	Baby's eye color. Eye color would not help a parent identify what is wrong with the baby.

Variable and Value	Marksmanship Coaching Example	Parenting Class (Sick Baby) Example
Allow for low relevancy of information source	Day of the week. Knowing what day of the week the scenario is taking place can only help identify the problem in that it might contribute to a shooter's mental state (later in the week, might be a bit more distracted because of the upcoming weekend). However, the coach could cope without knowing this information.	Time of the day. What time it is could explain the fussiness, but for the most part, most of the information is not dependent upon what time it is. The student can identify what is wrong with the baby without knowing this information.
Allow for medium relevancy of information source	Whether the shooter just recently completed classroom training on marksmanship. Knowing when the shooter completed classroom training on marksmanship may be an indicator of the shooter's experience and knowledge. A coach can still diagnose that the shooter is having problems with windage settings without this information, although having it would make it easier to rule out plausible diagnoses.	Whether the baby's sister has the flu. Knowing whether a sibling has the flu may be an indicator that the baby has the flu. However, a parent can still diagnose that the baby has the flu without knowing this information.
Allow for high relevancy of information source	A weapon's sight settings. In order to diagnose a shooter's problem, a coach would have to have access to the weapon's sight settings because it is a major factor in the shooting process.	Baby's temperature. One distinction between a cold and the flu is an onset of a high fever so this information is critical to identifying what is wrong with the baby.
Number of information sources		
Single	Student needs to make a diagnosis only looking at the target. The student is asked to identify a shooter's problem by simply examining the shots on the target.	Does not apply.
Multiple	Student makes a diagnosis based upon the target, the shooter, the weather information, and the data book. The student is asked to identify a shooter's problem using these pieces of information.	Student is asked to make a diagnosis based upon the baby's temperature, how much she has eaten that day, fussiness level, and severity of cough. The student is asked to identify a baby's sickness using these pieces of information.

Variable and Value	Marksmanship Coaching Example	Parenting Class (Sick Baby) Example
Prior knowledge required of task		
Low prior knowledge	Student is asked to provide a definition of trigger control.	Student is asked to provide what is a normal body temperature.
High prior knowledge	Student is asked to identify and fix the shooter's problem using multiple information sources.	Student is asked to identify the baby's sickness and prescribe the proper remedy using multiple information sources.
Prior knowledge of student		
Low prior knowledge	Assessment is taken at the beginning of the coaches course so it is assumed that the student has only a basic grasp of the coaching process. Presumably, the student will know the basic use of the information sources (e.g., data book) but not fully understand how they all relate to each other and can be used to identify a shooter's problem. The scenario might provide more information about the situation (e.g., state what the problem is, or partially identify the problem).	Assessment is given to students who have no experience with children. Presumably, the student will know little about babies. The parent might know that a temperature needs to be taken but might not know how high a "high temperature" should be. The scenario might provide more information about the situation (e.g., state what the problem is or provide help seeking options).
High prior knowledge	Assessment is taken at the end of the coaches course so it is assumed that the student has more knowledge about the coaching process. Assessment is designed to test whether a student learned the material therefore less guidance is provided on how to solve the problem.	Assessment is given at the end of the class so it is assumed that the student has more knowledge about parenting. Assessment is designed to test whether a student learned the material, therefore less guidance is provided on how to solve the problem.

Variable and Value	Marksmanship Coaching Example	Parenting Class (Sick Baby) Example
PROBLEM CHARACTERISTICS		
Type of task		
Execute	Not applicable.	Not applicable.
Fix	Student is told that the shooter has incorrect windage settings. The student is told why the shooter is having problems hitting the center of the target. Task is designed to assess whether a coach knows how to fix the windage settings to solve the problem.	Student is told that the baby has the flu. The student is told what is wrong with the baby. Task is designed to assess whether a parent knows what to do if the baby has the flu.
Change usual sequence	Student is told that the shooter has problems with hitting the center of the target and that one coach identified the problem by first looking at the target, then the range flag, and then the data book. Student is asked for a different way of solving the problem. Student can solve the problem a different way. For example, look at the data book first, then the target, then the range flag, and then back at the data book.	Student is told that the baby is sick and that one parent identifies what is wrong by first taking the temperature, then touching the baby's forehead, then checking the diaper, and then looking for other symptoms. Student is asked for a different way of solving the problem. Student can solve the problem a different way. First the parent can observe the baby's cough and congestion, then check the baby's diaper, then feel his forehead, then take his temperature, etc.
Improvise steps	Student is told that the shooter is missing his data book. Student needs to come up with a way to obtain the information normally found in the data book (target information, sight settings, etc.) to solve the problem.	Student is told that the thermometer is broken. Student needs to come up with a way to obtain the baby's temperature to solve the problem.

Variable and Value	Marksmanship Coaching Example	Parenting Class (Sick Baby) Example
Combination of tasks	Student is given the following scenario: shooter has incorrect windage settings but her data book keeps getting wet in the rain. One coach has been solving the problem by looking at the weapon and the target after each round. Student is asked to solve the problem using a different sequence of steps. This problem requires the student to improvise a way to obtain data book information, change the usual sequence the problem is solved, and to fix the problem once it is identified.	Student is given the following scenario: you are traveling with your baby, and she gets sick and has a fever. The parent needs to first bring down the fever and then figure out how to get the baby seen by a doctor. This problem requires a combination of tasks: fix the problem (bring down the fever) and improvise steps (get the baby seen by a doctor when normal pediatrician is out of town).
SOLUTION STRATEGY		
Steps		
Explicit course of action	Fix a shooter's windage settings by adding clicks to right. The fix to the problem has a prescribed procedure for remedy.	Help a baby with the flu using a sequence of actions. For example, student needs to first take a temperature, then check the diaper, and then call the doctor, etc.
Non-specified course of action	Have to diagnose a shooter who is missing his data book. There is no prescribed sequence of actions to diagnose a shooter who is missing his data book. Student will need to improvise a way to obtain the information.	How to determine whether a baby has a fever when the baby squirms too much so that the parent cannot take a temperature in the rectum. Student will need to improvise a way to obtain this information.
Problem subdivision		
Problem subdivision required	Does not apply—information is interrelated.	Does not apply—information is interrelated.
Problem subdivision not required	Student is told that the shooter has incorrect windage settings. Student knows that the source of the shooter's problem is her windage settings. There is no need to break it down further.	Student is told that the baby has the flu. Student knows the source of the baby's sickness, so there is no need to break it down further.

Variable and Value	Marksmanship Coaching Example	Parenting Class (Sick Baby) Example
Contingency planning		
Contingency planning required	Student has the hypothesis that the shooter's problem could either be with position or sight adjustment. The coach fixes the shooter's positioning. The student will need to watch the next shot to see if the applied solution was adequate. If not, the student will need to apply the back-up plan (adjust the sights).	This situation always requires a contingency plan because something could always go wrong. Contingency plans could include the knowledge of what to do if the fever shoots up to above 103° or what to do if the doctor cannot be called or if the baby cannot be soothed by rocking and holding.
Contingency planning not required	Fixing a shooter's incorrect windage settings. Student knows that adjusting the settings to compensate for the wind blowing in from the left will cause the shots to hit the center of the target.	Does not apply. Each solution always requires a backup plan because something might go wrong.
Help seeking		
Help seeking required	Does not apply.	Parent needs to consult a medical guide to help diagnose the baby's sickness. A scenario could present symptoms that are unfamiliar to the parent, therefore requiring consulting a medical guide to make use of the information.
Help seeking not required		Parent does not need to seek help to diagnose the baby's sickness.
Solution cognitive strategies		
Domain independent	Does not apply. Student cannot solve the shooter's problem with general algorithms.	Parent randomly tries a number of things to figure out what is wrong. Parent uses trial and error to see if he can identify why the baby is sick.
Domain dependent	Problem needs to be solved using specific marksmanship coaching strategies.	Problem needs to be solved using specific medical techniques.

Variable and Value	Marksmanship Coaching Example	Parenting Class (Sick Baby) Example
SOLUTION CHARACTERISTICS		
Solution space		
Convergent	The problem is that the shooter has incorrect windage settings. This problem has a known solution: to adjust the windage settings.	The baby has the flu. This problem has a known solution: call the doctor, attempt to bring the fever down, and keep the baby hydrated.
Divergent	Coach needs to be able to help a shooter whose data book is missing. This problem could be solved by a number of procedures: the coach can use the target, ask the shooter, or maintain a record of the shooter's performance.	The baby is crying, he doesn't have a fever or any sick symptoms, and the normal holding and rocking does not soothe him. This problem could be solved a number of ways: take a drive to distract the baby, give a pacifier, or call the doctor.
Solution correctness		
Multiple acceptable solutions	Coach needs to be able to help a shooter whose data book is missing. The solution to this problem can be solved in several ways (e.g., look at target or ask the shooter).	Parent needs to determine the baby's temperature but baby will not stop squirming. The solution to this problem can be solved in several ways (enlist help from someone else, put the thermometer in the armpit, etc.).
Partially acceptable solution	Coach needs to be able to help a shooter whose data book is missing. The student can demonstrate competence in making up for some missing information (e.g. by indicating that using target information could provide the missing information) but be unable to use the target information to make a diagnosis.	Parent knows that the temperature must be taken with a thermometer. However, the parent cannot use the information to make a diagnosis. Competence is demonstrated by the indication that a temperature must be taken, but the parent is unable to use the target information to make a diagnosis.

Variable and Value	Marksmanship Coaching Example	Parenting Class (Sick Baby) Example
Sub-solution contingencies		
Sequential	Shooting scenario is presented sequentially. While trying to diagnose the shooter, the student moves her positioning. The system updates the information sources to mirror this change.	Baby diagnosis scenario is presented sequentially. While trying to diagnose the baby, the parent gives her medicine. The system updates the information sources to mirror this change.
Non-sequential	The student is given the data book information, weather conditions, a picture of the shooter's position, and the shots of the target. Using these pieces of information, the student must identify the problem.	The parent is given all of the baby's symptoms and pertinent information. Using these pieces of information, the student must identify and solve the problem.